On Multiplexing Flows: Does it Hurt or Not?

Florin Ciucu
University of Warwick

Felix Poloczek
University of Warwick / TU Berlin

***Abstract*—This paper analyzes queueing behavior subject to multiplexing a stochastic process $M(n)$ of flows, and not a constant as conventionally assumed. By first considering the case when $M(n)$ is iid, it is shown that flows' multiplexing 'hurts' the queue size (i.e., the queue size increases in distribution). The simplicity of the iid case enables the quantification of the 'best' and 'worst' distributions of $M(n)$, i.e., minimizing/maximizing the queue size. The more general, and also realistic, case when $M(n)$ is Markov-modulated reveals an interesting behavior: flows' multiplexing 'hurts' but only when the multiplexed flows are sufficiently long. An important caveat raised by such observations is that the conventional approximation of $M(n)$ by a constant can be very misleading for queueing analysis.**

I. Introduction

Resource allocation is an old problem which perpetually reincarnates itself in resource sharing systems such as the telephone network, the Internet, or data centers. The first influential related treatment was performed by Erlang who essentially looked at the problem of dimensioning the telephone network. One of Erlang's main results was a formula for the computation of the blocking probability that some shared resource is occupied [20]; remarkably, amongst many applications, this formula has been used for nearly a century to dimension telephone networks.

Erlang's seminal work triggered the development of queueing theory, which has become an indispensable mathematical framework for the performance analysis of resource sharing based systems. Over almost a century, the exact approach to queueing theory (a.k.a. the classical approach) has been generalized to cover a broad class of networks, largely known by the product-form property (Baskett *et al.* [6], Kelly [25]). Besides its large scope, the class of product-form queueing networks is numerically tractable using convolution (Buzen [14]) or mean value analysis algorithms (Reiser and Lavenberg [34]).

Several alternative theories to queueing have been developed to avoid the general limitation of Poisson arrivals of product-form networks. One is the theory of effective bandwidth (Kelly [26], Mazumdar [33]), which relies on large deviation techniques and provides a rather straightforward analysis of multiplexing regimes for a broad class of arrival processes. An extension of the effective bandwidth theory, which can additionally deal with many scheduling algorithms and especially multi-queue scenarios, is the stochastic network calculus (Chang [15], Jiang and Liu [24], Ciucu and Schmitt [18]). These conceivably attractive theories provide, however, queueing metric results in terms of either exact asymptotics or probabilistic bounds. While the relevance of asymptotics and (conceivably loose) bounds is often questioned (Abate *et al.* [1], Choudhury *et al.* [17], Shroff and Schwartz [40]), other advanced techniques yield much refined results (Duffield [19], Liu *et al.* [30], Chang [15], Mazumdar [33]) at the expense however of a more involved analysis.

The common challenge faced by queueing approaches, when modelling some unpredictable resource sharing based system, is capturing the system's inherent randomness. For instance, in the context of a network router, the high variability inherent to packet flows is captured by conventional queueing models with probability distributions (e.g., of the packets' inter-arrival times and sizes). Network calculus models use instead envelope functions, which enforce either deterministic or probabilistic bounds on the amounts of packets over time intervals. A very recent alternative to classical queueing theory uses deterministic models satisfying the implications of probability laws characteristic to the packets flows (e.g., the Law of the Iterated Logarithm, see Bertismas *et al.* [8]).

While capturing randomness is essential in modelling, different randomness models can lead to very different (and possibly bogus) insights on actual system behavior. Consider for instance a simple example of a router with capacity C which is being modelled by the classic M/M/1 queue: packets arrive as a Poisson process with rate λ, and their sizes are exponentially distributed with average $1/\mu$. Under the stability condition $\lambda/(\mu C) < 1$, the packets' average delay is

$$E\Big[delay\Big] = \frac{1}{\mu C - \lambda} \ . \tag{1}$$

Consider next the much simpler averaged-out D/D/1 model, in which the interarrival times are constant (i.e., equal to $1/\lambda$) and packet sizes are constant as well (i.e., equal to $1/\mu$). Under the same stability condition, the packets' average delay becomes

$$E\Big[delay\Big] = \frac{1}{\mu C} \ . \tag{2}$$

Note the different quantitative results predicted by the two models, with the observation that the 'more-random' one predicts higher delays. Such stochastic ordering properties, formalizing the manifestation of the folk principle that "*determinism minimizes the queue*", have been studied in the context of queueing systems (see the related work section) and even for risk management (see, e.g., Asmussen *et al.* [3]).

Let us consider a more complex queueing model subject to flows' multiplexing and which explicitly accounts for the number of parallel flows, denoted throughout by $M(n)$. While there is an overwhelming work on *static queues* whereby $M(n)$ is a constant, much less is known on *dynamic queues*

whereby $M(n)$ is a stochastic process[1]. Moreover, since communication networks are more accurately modelled by dynamic queues (e.g., the number of parallel flows traversing an Internet router *is* a stochastic process) the goal of this paper is to provide an analytical understanding on the role of randomness in $M(n)$ on the queue size (e.g., How fast does it grow?). In particular, the paper attempts to provide insights into the illustrative question "Multiplexing Flows: Does it Hurt or Not?", rephrased as "What is the *joint impact* of stochastic models, for both $M(n)$ and the flows' themselves, on the queue size?".

To answer such a fundamental question we consider two randomness models. One is subject to strong iid (independent and identically distributed) assumptions, enabling a tractable analytical study on the impact of various distributions of $M(n)$ on the queue size. The second more realistic case is when $M(n)$, and also the flows, have a Markov structure. While stochastic bounds on the queue size can also be derived, as in the iid case, they are expressed in terms of eigen-values/vectors hampering an explicit analytical investigation; for this reason, numerical evaluations will be invoked.

By using convexity arguments, the simplicity of the iid case enables showing that the best-case distribution from the perspective of the queue size is the intuitively obvious *constant distribution*, extending thus the folk principle that "determinism minimizes the queues" from static to dynamic queues. The second extremal property concerns the corresponding worst-case distribution, i.e., which law of $M(n)$ maximizes the queue size? It is shown that this is a *bimodal distribution*, with mass on the extremes of $M(n)$'s range and therefore maximizing all the moments. This result also agrees with parallel results from static queues concerning extremal properties of bimodal distributions (see Section II-C). Another immediate result is that strong conditions on ordering distributions are needed, in contrast to parallel results from M/G/k queues. The perhaps most fundamental insight is that the above folk principle can fail, in the more realistic case when $M(n)$ is Markov-modulated. Concretely, we find that there is a phase transition in the flows' average lifetimes at which dynamic queue models yield (stochastically) larger queues than the corresponding (normalized) static queue models.

These overall insights raise the important caveat that approximating (realistic) dynamic queues by static queues (i.e., replacing the stochastic process $M(n)$ by its mean $E[M(n)]$) can yield very misleading results, which can either overestimate or underestimate the 'true' results.

The rest of the paper is structured as follows. First we overview related work. In Section III we treat dynamic queues under iid multiplexing, and in Section IV under more realistic Markovian assumptions. Section V summarizes the paper.

[1]We use the terminologies *static queue* when the number of parallel flows is deterministic and *dynamic queue* when the number of flows is random. While not standard and perhaps confusing, the terminology is preferred as a convenient shorthand.

II. Related Work

Here we overview previous work related to the main topics of this paper, i.e., 1) the relevance of studying dynamic queues, 2) stochastic orderings concerning queueing metrics, and 3) extremal distributions for minimizing/maximizing queues.

A. Dynamic Queues and Analytical Approaches

The importance of accounting for the elastic nature of Internet traffic, determined by a dynamic or random number of parallel flows, has been recognized in the context of bandwidth sharing. Massoulié and Roberts showed that randomness in the number of parallel flows can have unpredictable consequences on the throughput of long-lived flows, irrespective of the assigned weights to the parallel flows [32]. In a similar setting, Bonald and Massoulié demonstrated that network stability is insensitive to a broad range of fair allocations [10], generalizing a result of de Veciana *et al.* for weighted max-min fairness [43]. A more recent study of Liu *et al.* showed that stability is actually sensitive to the settings of α-fairness, in networks with non-convex and time-varying rate regions [29], generalizing an earlier result of Bonald and Proutière [11]. Another notable insensitivity result is that in dynamic scenarios with flows arriving as a Poisson process, the first moments of the number of flows and the flows' throughput do not depend on the flow size distribution or on the properties of the flows' arrivals (Fred *et al.* [20]).

A general way to model randomness in the number of flows is through a queue with bulk arrivals, i.e., the $G^{[M]}/G/1$ queue, whereby customers arrive in batches of random size M according to a renewal process, and customers have some service time distribution. In the case of Poisson renewals, exact solutions exist for various queueing metrics (e.g., Laplace transforms for waiting times) and various scheduling of the batches: FIFO (Burke [12]), with priorities (Takagi and Takahashi [42]), or PS (Bansal [5]); for more general renewals solutions are given numerically (Schleyer [38]) or in terms of bounds (Yao *et al.* [47]). For an excellent treatment of queues with bulk arrivals see Chaudhry and Templeton [16]. Our contribution herein is to analyze very general distributions (subject to a finite moment generating function (MGF)).

Other analytical approaches address queueing models with fluid arrivals. For instance, the classical Anick-Mitra-Sondhi model [2], with a fixed number of flows producing arrivals at some rates according to the states of Markov On-Off processes, can be regarded as a queue with a binomial number of flows. Queueing in related fluid models can be analyzed exactly in terms of spectral representations, at a cost of high computational complexity due to a combinatorial explosion in the number of states [41]. The advantage of our approach is that it provides *simple* (convex) upper and lower bounds on queueing metrics, which further permit the immediate analysis of extremal properties.

B. Stochastic Orderings

Stochastic orderings, setting partial orders for queueing metrics, were previously addressed in static scenarios. An

elementary example on the role of the variability of underlying distributions was just illustrated in Eqs. (1) and (2). More generally, in M/G/k queues, the average delay was shown to be an increasing function of the variance of the service time distribution (see Whitt [44], [45]). Extensions of this monotonicity property were considered by Asmussen and O'Cinneide in [4] for Markov-modulated M/G/1 queues. For single queues with Markov-modulated Poisson processes, and under some monotonicity assumptions on the generator of a Markov chain modulating the intensity, Bäuerle and Rolski [7] proved that the queues increase by scaling down the generator. In the case of networks with Poisson arrivals, it was shown that exponential packet sizes yield smaller delays than averaged-out sizes but not in full generality (for a counterexample see Harchol-Balter and Wolfe [22]). When the arrivals are not Poisson however, the monotonicity property fails in some cases even for single queues (see, e.g., Ross [37]).

This paper shows that the monotonicity of the variance alone of the number of flows $M(n)$ is *not* sufficient to infer stochastic orderings on the queue size; instead, a sufficient condition is given by the monotonicity of the MGF. In the light of related work, our result thus indicates that queuing metrics are much more sensitive to the variability of the number of flows than of the flows themselves; this claim is further supported by the emphasized sensitivity of dynamic queues to peak rather than average-values.

C. Extremal Distributions

A "folk theorem" in queueing theory states that, when the average inter-arrival (service) time is fixed, the constant inter-arrival (service) time distribution *minimizes* queueing metrics such as average waiting time. This result was proven for renewal processes (see Rogozin [35]) and also for more general arrival processes with exponential service times (see Hajek [21] and Humblet [23]). A related variant of the underlying intuitive principle that "determinism minimizes the waiting" is that round-robin server assignment outperforms random server assignment (see Makowski and Philips [31]).

In turn, bimodal distributions maximize queue lengths in GI/M/1 queues (Whitt [46]), in G/M/1 queues with bulk arrivals (Lee and Tsitsiklis [28]), and more recently in queues with bulk arrivals and finite buffers (Bušić *et al.* [13]). We will show that these extremal properties characteristic to static queues extend to dynamic queues as well.

III. IID Multiplexing

We first consider multiplexing under strong iid assumptions of the flows. This simplified case enables an analytical study on the impact of the distribution of the number of parallel flows on the queue size. For the more realistic Markov-modulated multiplexing case, which is only amenable to a numerical study, see the next section.

We consider the single-queue scenario from Figure 1. The queue has an infinite sized buffer, whereas the server has a constant capacity C and serves the arrivals in a work-conserving manner.

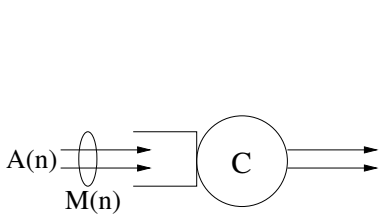


Fig. 1. A server with constant rate C serving a single queue with input $A(n)$ consisting of $M(n)$ parallel flows.

After introducing the arrival model, we will derive upper and lower bounds on the queue size, and then discuss on extremal distributions of $M(n)$ relative to achievable queue sizes; the obtained analytical insights will be finally complemented by some illustrative numerical results.

A. Arrival Model

The time model is discrete. The number of parallel flows active at time n is represented by a stationary stochastic process $M(n)$. The cumulative arrival process $A(n)$, counting the number of data units (e.g., packets) over the time interval $[0, n]$ is defined recursively as

$$A(n) = A(n-1) + \sum_{i=1}^{M(n)} a_i(n) \ , \tag{3}$$

with the initial condition $A(0) = 0$. The instantaneous arrival process at time n is represented by the random vector $\mathbf{a}(n) = (a_1(n), a_2(n), \dots)$. When clear from the context, we will refer to the elements of $M(n)$ by M, and to the elements of $\mathbf{a}(n)$ simply by a.

For some $\theta > 0$, we assume that the moment generating functions (MGFs)

$$\phi_a(\theta) := E\left[e^{\theta a}\right] \text{ and } \phi_M(\theta) := E\left[e^{\theta M}\right]$$

are finite. Moreover, for the sake of simplicity we assume that the elements of $\mathbf{a}(n)$ and $M(n)$ are each iid (independent and identically distributed), and jointly independent.

B. The Queue Distribution

Since the increment process $A(n) - A(n-1)$ is reversible, the stationary queue length Q can be written as

$$Q = \sup_{n \geq 0} \{A(n) - Cn\} \ .$$

The next theorem provides upper and lower bounds for the distribution of Q .

Theorem 1. *(Q's Distribution, iid-case) Consider the arrival process from Eq. (3) and assume that the elements of* $\mathbf{A}$ *are iid with MGF* $\phi_a(\theta)$*, and the elements of* $\mathbf{M}$ *are iid with MGF* $\phi_M(\theta)$*; also,* $\mathbf{A}$ *and* $\mathbf{M}$ *are independent. Consider a queue with service rate* C *and let*

$$\theta := \sup\left\{\theta' \geq 0 : \phi_M\left(\log \phi_a(\theta')\right) = \phi_C(\theta')\right\} \ . \tag{4}$$

Then we have the upper bound for all $x \geq 0$

$$\mathbb{P}\Big(Q \geq x\Big) \leq e^{-\theta x} \ . \tag{5}$$

If in addition there exists the constants a_{max} and N_{max} such that $a_1(1) \leq a_{max}$ almost surely (a.s.), $M(1) \leq N_{max}$ a.s., and $N_{max}a_{max} > C$, then we have the lower bound for all $x \geq 0$

$$\mathbb{P}\Big(Q \geq x\Big) \geq e^{-\theta(N_{max}a_{max}-C)}e^{-\theta x} \ .$$

The upper and lower bounds are asymptotically *exact* (i.e., the following limit $\lim_{x\to\infty} \frac{1}{x}\log \mathbb{P}\left(Q > x\right) = \theta$ holds) since the two exponential bounds have the same decay rate θ. We remark that the theorem immediately extends to the case of a queue with random instantaneous capacities $(C(1), C(2), \dots)$, if these are iid; the only modification is that $\phi_C(\theta)$ in Eq. (4) is to be replaced by $\phi_{C(1)}(\theta)$. In the theorem, we do not explicitly impose the stability condition $\phi'_a(0)\phi'_M(0) < C$. Unless this is true then $\theta = 0$ in Eq. (4). Also, for the lower bound, the condition $N_{\max}a_{\max} > C$ avoids the trivial situation of no queueing.

To prove the upper bound we apply Kingman's technique for GI/GI/1 queues based on an exponential martingale [27]. To prove the lower bound we rely on some additional ideas from Ross [36] and Chang [15].

Proof. Let $x \geq 0$. With $\theta > 0$ as in the theorem we construct the random process

$$X_n = e^{\theta(A(n)-Cn)}$$

for all $n \geq 0$. Let also the associated filtration of σ-algebras $\mathcal{F}_n = \sigma(\mathbf{a}(1), \dots, \mathbf{a}(n), M(1), \dots, M(n))$, where $\mathbf{a}(n)$'s denote the vectors $(a_1(n) \quad a_2(n) \quad \cdots)$.

The key to the proof is to show that X_n is a martingale. For some $n \geq 1$ we can write for the conditional expectation

$$\begin{aligned} E\left[X_n \mid \mathcal{F}_{n-1}\right] &= E\left[X_{n-1}e^{\theta\left(\sum_{i=1}^{M(n)} a_i(n)-C\right)} \mid \mathcal{F}_{n-1}\right] \\ &= X_{n-1}E\left[e^{\theta\left(\sum_{i=1}^{M(n)} a_i(n)-C\right)}\right] \ , \end{aligned}$$

using that X_{n-1} is $\mathcal{F}_{n-1}$-measurable and the independence assumptions on $\mathbf{A}$ and $\mathbf{M}$. Further conditioning on $M(n)$ we can compute the last expectation

$$\begin{aligned} E\left[e^{\theta\sum_{i=1}^{M(n)} a_i(n)}\right] &= \sum_{m\geq 0} \phi_a(\theta)^n \mathbb{P}\left(M(n) = m\right) \\ &= \phi_M\left(\log \phi_a(\theta)\right) \ , \end{aligned}$$

after using the independence properties again. With this we can continue above

$$E\left[X_n \mid \mathcal{F}_{n-1}\right] = X_{n-1}\phi_C(-\theta)\phi_M\left(\log\phi_a(\theta)\right) = X_{n-1} \ ,$$

using the definition of θ. Therefore the sequence X_n is a martingale (relative to $\mathcal{F}_n$).

The second part of the proof (for the upper bound) roughly reproduces Doob's inequality; let us define

$$T = \inf\left\{n \geq 0 : A(n) - Cn \geq x\right\}$$

as the first passage time to exit $[0, x]$. Note that T is a stopping time relative to $\mathcal{F}_n$, i.e., $\{\omega : T_\omega \leq n\} \in \mathcal{F}_n$ for all $n \geq 0$.

Let $n \geq 0$. Then $T \wedge n := \min\{T, n\}$ is a bounded stopping time and according to the optional sampling theorem (see Billingsley [9], p. 466) applied to the martingale X_n we have

$$\begin{aligned} E\left[X_0\right] &= E\left[X_{T\wedge n}\right] \geq E\left[X_{T\wedge n}I_{\{T\leq n\}}\right] \\ &\geq e^{\theta x}\mathbb{P}\left(T \leq n\right) \ , \end{aligned} \tag{6}$$

where $I_{\{\cdot\}}$ denotes the indicator function; in the last line we used the definition of T. Letting $n \to \infty$ we obtain

$$E\left[X_0\right] \geq e^{\theta x}\mathbb{P}\left(T < \infty\right) \ .$$

Finally, using

$$P\left(T < \infty\right) = \mathbb{P}\left(\sup_{n\geq 0}\left\{A(n) - Cn\right\} \geq x\right) \tag{7}$$

and $E\left[X_0\right] = 1$, we immediately get the upper bound from the theorem.

To prove the lower bound we further let $y \geq 0$ and denote

$$T_y = \min\left\{T, \inf\left\{n \geq 0 : A(n) - Cn \leq -y\right\}\right\}$$

as the first time to exit the interval $[-y, x]$. Note that T_y is a finite stopping time relative to $\mathcal{F}_n$. By the optional stopping theorem, the process $(X_{T_y\wedge n})_n$ is a martingale, which is bounded and hence uniformly integrable. Thus, $X_{T_y\wedge n} \to X_{T_y}$ a.s. and in L^1, and we have

$$\begin{aligned} &E\left[X_0\right] = E\left[X_{T_y\wedge 0}\right] = E\left[X_{T_y}\right] \\ &= E\left[X_{T_y} \mid A\left(T_y\right) \geq CT_y + x\right] P\left(A\left(T_y\right) \geq CT_y + x\right) \\ &+E\left[X_{T_y} \mid A\left(T_y\right) \leq CT_y - y\right] P\left(A\left(T_y\right) \leq CT_y - y\right) \ . \end{aligned} \tag{8}$$

Note further the implications of events

$$\begin{aligned} \left\{A\left(T_y\right) \geq CT_y + x\right\} &\Rightarrow \left\{T_y = T\right\} \\ &\Rightarrow \left\{A\left(T_y - 1\right) < C\left(T_y - 1\right) + x\right\} \\ &\Rightarrow \left\{A\left(T_y\right) \leq CT_y + N_{\max}a_{\max} - C + x\right\} \ , \end{aligned}$$

where we used the definition of T and the bounding constants from the theorem. We can thus bound the previous sum as

$$E\left[X_0\right] \leq e^{\theta(N_{\max}a_{\max}-C+x)}P\left(A\left(T_y\right) \geq CT_y + x\right) + e^{-\theta y} \ .$$

Letting $y \to \infty$ yields

$$E\left[X_0\right] \leq e^{\theta(N_{\max}a_{\max}-C+x)}P\left(T < \infty\right) \ .$$

The lower bound from the theorem follows immediately from Eq. (7) and $E\left[X_0\right] = 1$, which completes the proof. □

C. Extremal Distributions

Given the bounds from Theorem 1, we can identify the best/worst-case distributions for $M(n)$ which minimize/maximize the queue size. Then we discuss on conditions under which a particular distribution is 'better' or 'worse' than another.

To formalize the underlying stochastic ordering, and thus the meaning of 'better' and 'worse', we say that a queue Q_1 is smaller than another queue Q_2 if the corresponding decay rates θ_1 and θ_2 (e.g., defined in Eq. (4)) satisfy

$$\theta_1 \leq \theta_2 \ .$$

1) Best-Case Distribution: First we briefly show the intuitive result that the best-case distribution of M is the constant one. What is more interesting is that neither of the distributions of M and a dominates the other, when jointly accounting for both.

Given the iid assumption, Jensen's inequality applied to the exponential function (i.e., $e^{\theta E[X]} \leq E\left[e^{\theta X}\right]$ for some r.v. X) yields that

$$\phi_{E[M]}\left(\log \phi_a(\theta)\right) \leq \phi_M\left(\log \phi_a(\theta)\right) \ .$$

The left-hand side corresponds to the composition of MGFs from the definition of θ from Eq. (4) when there is no randomness in the number of parallel flows, i.e., when the elements of $M(n)$ are equal to a single constant. In turn, the right-hand side accounts for randomness in $M(n)$. Because of the inequality above, it follows that the value of θ from Eq. (4) decreases when accounting for randomness, which further means that the queue increases correspondingly. The best-distribution is thus the constant, which in particular minimizes all the moments.

Finally, we point out the interesting fact that none of the randomness in the number of parallel flows, or at the flow level, dominates the other. That is because there is no general ordering between the terms

$$\phi_{E[M]}\left(\log \phi_a(\theta)\right) \text{ and } \phi_M\left(\log \phi_{E[a]}(\theta)\right) \ .$$

Indeed, using Jensen's inequality, the left term is the smallest when a is non-random (i.e., $a = E[a]$) and M is random. In turn, the left term is the largest when M is non-random (i.e., $M = E[M]$) and a is random. This fundamental lack of monotonicity suggests that, even for the purpose of deriving bounds on the queue size distribution, both the randomness in the number of flows and at the flow level must be jointly accounted for. In other words, simplifying the queueing model by averaging-out either M or a can lend itself to bogus results.

2) Worst-Case Distribution: According to Theorem 1, the problem of determining the distribution of M which maximizes the queue reduces to solving for

$$\underset{M,\ \text{fixed}\ E[M]}{\operatorname{argmax}} \ E\left[e^{\theta M}\right] \ , \tag{9}$$

for all $\theta > 0$. The next Lemma gives the solution.

Lemma 1. *(*WORST-CASE DISTRIBUTION*) Assuming that M has the support $\{0, 1, \ldots, M_{max}\}$, the solution of Eq. (9) is the bimodal distribution with*

$$\pi_0 = 1 - \frac{E[M]}{M_{max}} \text{ and } \pi_{M_{max}} = \frac{E[M]}{M_{max}} \ .$$

PROOF. Assume that there exists $0 < i < M_{\max}$ such that $\pi_i := \mathbb{P}(M = i) > 0$. Denoting $x = \frac{M_{\max} - i}{M_{\max}} \pi_i$, let us observe that

$$\pi_0 + \pi_i e^{\theta i} + \pi_m e^{\theta M_{\max}} \leq \pi_0 + x + \left(\pi_{M_{\max}} + \pi_i - x\right) e^{\theta M_{\max}} \ . \tag{10}$$

Indeed, showing this inequality reduces to showing that the function

$$f(i) := \frac{e^{\theta M_{\max}} - e^{\theta i}}{M_{\max} - i}$$

is monotonically increasing over $i \in \{0, 1, \ldots, M_{\max} - 1\}$. This can be shown immediately by extending $f(\cdot)$ to continuous time, differentiating, and using the inequality $e^z \geq z + 1$ for $z \geq 0$.

Therefore, Eq. (10) shows that a 'worse' distribution can be obtained by appropriately spreading the distribution mass to the extremes. Note that the new distribution retains the average value $E[M]$ since

$$i\pi_i + m\pi_{M_{\max}} = M_{\max}\left(\pi_{M_{\max}} + \pi_i - x\right) \ .$$

The proof is complete by repeatedly spreading the mass, as in Eq. (10), for all $0 < i < M_{\max}$ for which $\pi_i > 0$. □

We note that the bimodal distribution was found to attain the maximum over a partial order set according to convex ordering (see Shaked and Shanthikumar [39], Theorem 3.A.24, p. 125); in our case, the ordering is restricted to MGFs only.

D. Ordering Distributions

The constant best-case distribution and the bimodal worst-case distribution identified earlier are clearly unrealistic from a practical point of view. It is thus of interest to analyze the relationship between different (and more realistic) distributions from the point of view of being 'better' or 'worse'.

Following the presented arguments, an immediate sufficient condition for a distribution M_1 to be 'better' than a distribution M_2 (subject to the condition $E[M_1] = E[M_2]$) is an ordering on the MGFs, i.e.,

$$E\left[e^{\theta M_1}\right] \leq E\left[e^{\theta M_2}\right] \ , \tag{11}$$

for all $\theta > 0$. This can be seen from the construction of the optimal θ from, e.g., Eq. (4) in Theorem 1.

The condition from Eq. (11) is clearly strong as it implicitly involves all the moments of M_1 and M_2. In the light of the discussion from Section II-B that an ordering on the variance (of packet distributions) is sufficient for ordering the queue sizes in M/G/k queues, we point out that a similar condition on the variance is not sufficient in the current context (mainly due to non-Poisson input). To quickly illustrate this negative fact, by counterexamples, let $C = 3$, M_1 the Uniform distribution with support $\{0, 1, 2, 3, 4\}$ and M_2 having the same support, the same average $E[M_1] = E[M_2] = 2$, and the mass $\pi_1 = 0.5$, $\pi_2 = 0.25$, and $\pi_4 = 0.25$. One can show that $Var[M_1] > Var[M_2]$ and

$$\sup\left\{\theta : e^{\theta C} = E\left[e^{\theta M_1}\right]\right\} > \sup\left\{\theta : e^{\theta C} = E\left[e^{\theta M_2}\right]\right\} \ , \tag{12}$$

i.e., M_1 is 'better' than M_2.

In turn, by changing the mass of M_2 to $\pi_1 = 0.5$ and $\pi_4 = 0.5$, one can show that $Var[M_1] < Var[M_2]$ but M_1 is 'worse' than M_2. To conclude, the variance alone of M is not a sufficient indicator for ordering the queues. Moreover, in the light of the above counterexamples, it is conceivable that the sufficient condition from Eq. (11), which imposes an ordering on the MGFs, is also necessary.

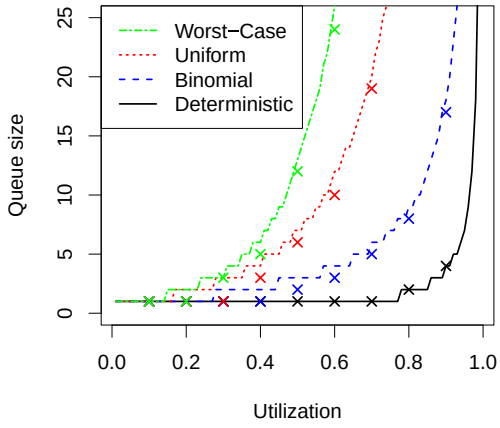


Fig. 2. Impact of several distributions for the number of parallel flows M on the queue size. Analytical bounds are depicted with lines, whereas corresponding simulation results are depicted with the '$\times$' symbol.

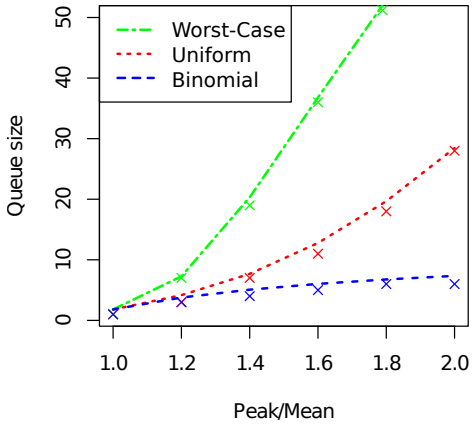


Fig. 3. Impact of several distributions for the number of parallel flows M on the queue size, depending on the peak-to-mean ratio.

E. Numerical Results

We now provide numerical evidence on the discrepancy between static and dynamic queues, by varying the distribution of the number of parallel flows M and also the corresponding peak-to-mean ratios.

To keep the analysis concise, we consider a homogenous scenario in which the elements of a are Bernoulli random variables taking the values 0 and 1 with probabilities $1-p$ and p, respectively. Figure 2 illustrates the queue size x, for a fixed violation probability $\varepsilon = 10^{-3}$, and as a function of the utilization factor; the other parameters are $E[M] = 10$, $M_{\max} = 20$, $C = 9$, and p is scaled accordingly for each utilization value. The worst-case distribution is the one from Lemma 1. The figure indicates that the impact of M's distribution on the queue size can be substantial (e.g., as large as many orders of magnitude). Moreover, simulation results (depicted with the '$\times$' symbol, for each distribution) indicate that our analytical bounds are quite tight.

In Figure 3 we illustrate the impact of several distributions on the queue size, especially when varying the peak-to-mean ratio (the same parameters are used as in Figure 2, except for scaling the peak and fixing the utilization to 75%). The figure provides strong evidence that approximating dynamic by static queues can be arbitrarily misleading for queueing metrics, even for moderate values of the peak-to-mean ratio.

As a side remark, the obtained results uncover several fundamental similarities and differences amongst the concepts of capacity when defined in 1) information theory (e.g., as the channel capacity), 2) static, and 3) dynamic queues (e.g., as the required capacity to guarantee some queueing constraints). All three corresponding maximal capacities are attained by the intuitively obvious constant distribution, which in particular has zero entropy. In turn, while the minimal channel capacity is attained by the uniform distribution (which maximizes the entropy), the two queueing minimal capacities are attained by bimodal distributions; this conceptual difference stems from the different scalar measures of a distribution used in information theory (i.e., the entropy) and queues (i.e., moments accounting for actual values).

IV. Markov-Modulated Multiplexing (MMM)

In this section we consider the Markov-Modulated Multiplexing (MMM) case, i.e., $M(n)$ is modulated by a Markov process. While MMM is more realistic than iid multiplexing, the implicit nature of the obtained stochastic bounds only allows for qualitative insights on the behavior of dynamic queues using numerical results.

A. Arrival Model

To model MMM we consider a number of $M_{\max}$ Markov-Modulated sources. For each source, transmissions are modulated by a Markov chain with state space $\mathcal{S} = \{0, 1, \text{IA}\}$ (see Figure 4).

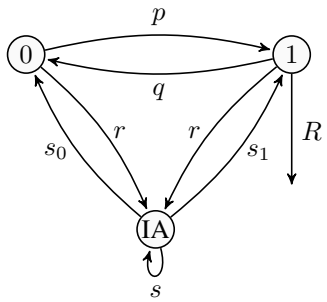


Fig. 4. A Markov process modulating the arrival process of a source

The upper two states correspond to a typical Markov-Modulated On-Off (MMOO) source which is idle while in state '0' and transmits at constant rate R while in state '1'. The extra state 'IA' models the situation that the MMOO source may be inactive. The difference between the states '0' and 'IA' is that $r << q$, i.e., it is much less likely for the source to enter the inactive state than the idle state. From the inactive state, the

source reactivates according to the (conditional) steady-state probability vector of the MMOO source, i.e.,

$$\pi_{\mathbf{act}} = \left(\frac{q}{p+q}, \frac{p}{p+q}\right) ,$$

such that $s_0 = \frac{q}{p+q}(1-s)$ and $s_1 = \frac{p}{p+q}(1-s)$. The transition matrix of the entire Markov chain is

$$T = \begin{pmatrix} (1-p)(1-r) & p(1-r) & r \\ q(1-r) & (1-q)(1-r) & r \\ \frac{q}{p+q}(1-s) & \frac{p}{p+q}(1-s) & s \end{pmatrix} . \tag{13}$$

To summarize, the number of active sources (i.e., parallel flows) is a (Markov) process $M(n)$ with support $\{0, 1, \ldots, M_{\max}\}$. The fundamental difference from the iid multiplexing model from Eq. (3) is that MMM allows for the dynamic multiplexing of bursty sources (e.g., MMOO processes). In particular, we point out that the model from Eq. (3) cannot be simply extended to bursty sources by relaxing the condition that the elements of $\mathbf{A}$ are iid; for instance, in the case of MMOO sources in Eq. (3), their Markovian structure would be ambiguous due to dynamically changing $M(n)$. On the other hand, the proposed MMM model restricts the distribution of $M(n)$ to a binomial, albeit the dynamical structure (i.e., driven by an implicit Markov chain) of $M(n)$ is captured.

B. The Queue Distribution

Let $(a_i(n))_n$, $i \in \{1, \ldots, M_{max}\}$, denote M_{max} independent copies of Markov-Modulated sources as in Figure 4. Then, the (cumulative) arrival process $A(n)$ is given by

$$A(n) = A(n-1) + \sum_{i=1}^{M_{max}} f(a_i(n)) , \tag{14}$$

where

$$f(x) := \begin{cases} 1 & x = 1 \\ 0 & x \in \{0, \mathrm{IA}\} \end{cases} .$$

It is easy to check that the stationary distribution of each source is given by the probability vector

$$\pi = \left(\frac{q(1-s)}{(p+q)(r+1-s)}, \frac{p(1-s)}{(p+q)(r+1-s)}, \frac{r}{r+1-s}\right) .$$

Further, the balance equations

$$\pi_i T_{i,j} = \pi_j T_{j,i} , \quad i, j \in \mathcal{S}$$

hold so that the sources $a_i(n)$, and hence the increment process $A(n) - A(n-1)$, are reversible. Consequently, the stationary queue length Q can be written as

$$Q = \sup_{n \geq 0} \{A(n) - Cn\} .$$

The key tool to bound Q's distribution is the following exponential column-transform of the transition matrix.

Definition 1. *For T from Eq. (13) and $\theta \geq 0$ define the exponentially transformed matrix T^θ as:*

$$T^\theta_{i,j} := T_{i,j} e^{\theta f(j)} , \quad i, j \in \mathcal{S} ,$$

i.e., the second column of T is multiplied with the factor $e^{\theta f(j)}$. Further, let $\lambda(\theta)$ denote the maximal positive eigenvalue and ν a corresponding positive eigenvector.

As T^θ is a nonnegative matrix, by the Perron-Frobenius-Theorem, $\lambda(\theta)$ equals to the spectral radius and an eigenvector ν with positive entries exists. The next theorem provides upper and lower bounds on Q's distribution.

Theorem 2. *(Q's Distribution, MMM-case) Consider the arrival model from Eq. (14) and a constant server capacity $C > 0$. Let*

$$\theta := \sup\{\theta \geq 0 : \lambda(\theta) = \phi_{CM^{-1}_{max}}(\theta)\} ,$$

then the following bounds on the backlog hold:

$$\mathbb{P}(Q \geq x) \leq H_u e^{-\theta x}$$
$$\mathbb{P}(Q \geq x) \geq H_l e^{-\theta x} ,$$

where

$$H_u = \frac{\left(\pi_0\nu_0 + \pi_1\nu_1 + \pi_{\mathrm{IA}}\nu_{\mathrm{IA}}\right)^{M_{max}}}{\nu_1^{\lceil CR^{-1}\rceil} + \min\{\nu_{\mathbf{0}}, \nu_{\mathrm{IA}}\}^{M_{max} - \lceil CR^{-1}\rceil}} \quad \textit{and}$$
$$H_l = \frac{\left(\pi_0\nu_0 + \pi_1\nu_1 + \pi_{\mathrm{IA}}\nu_{\mathrm{IA}}\right)^{M_{max}}}{\max_s \nu_s^{M_{max}} e^{\theta(RM_{max} - C)}} .$$

Note that the definition of θ resembles the one from Theorem 1 with the only difference that the MGF is replaced by the spectral radius. We also note that $\theta = 0$ when the queue is not stable, and that the upper and lower bounds are asymptotically exact since they have the same decay rate θ.

Proof. For $0 \leq i \leq M_{max}$ consider the process

$$X^i_n := \nu_{a_i(n)} e^{\theta\left(\sum_{k=1}^n f(a_i(k)) - CM^{-1}_{max} n\right)} .$$

In Duffield (see [19]) it is shown that X^i_n is a martingale. By the independence assumption on the M_{max} arrivals the product

$$X_n := \prod_{i=1}^{M_{max}} X^i_n = \prod_{i=1}^{M_{max}} \nu_{a_i(n)} e^{\theta(A(n) - Cn)}$$

is a martingale as well. Now similarly as in the proof of Theorem 1 define the stopping time

$$T = \inf\{n \geq 0 : A(n) - Cn \geq x\}$$

and then apply the optional sampling theorem to $T \wedge n$, implying that

$$\mathbb{E}[X_0] = \mathbb{E}[X_{T\wedge n}] \geq \mathbb{E}[X_{T\wedge n} I_{\{T \leq n\}}]$$
$$\geq e^{\theta x} \mathbb{E}[\prod_{i=1}^{M_{max}} \nu_{a_i(T)} I_{\{T \leq n\}}] .$$

A critical observation is that since T is the *first* point where $A(n) - nC \geq x$, the T'th increment is positive, i.e., at time

T at least $\lceil CR^{-1} \rceil$ chains are transmitting. Therefore:

$$\prod_{i=1}^{M_{max}} \nu_{a_i(T)} \geq \nu_1^{\lceil CR^{-1} \rceil} + \min\{\nu_{\mathbf{0}}, \nu_{\text{IA}}\}^{M_{max} - \lceil CR^{-1} \rceil}$$
$$= \frac{\mathbb{E}[X_0]}{H_u}$$

The upper bound then follows as in the proof of Theorem 1 by letting $n \to \infty$ and observing that

$$\mathbb{P}(Q \geq x) = \mathbb{P}(T < \infty) \ .$$

For the lower bound, define the stopping time

$$T_y = \min\{T, \inf\{n \geq 0 : A(n) - Cn \leq -y\}\}$$

for some $y \geq 0$. Using the same arguments as in the proof of Theorem 1 we have

$$\begin{aligned}\mathbb{E}[X_0] =& \mathbb{E}[X_{T_y} \mid A(T_y) - CT_y \geq x]\mathbb{P}(A(T_y) - CT_y \geq x) \\ &+ \mathbb{E}[X_{T_y} \mid A(T_y) - CT_y \leq -y]\mathbb{P}(A(T_y) - CT_y \leq -y) \\ &\leq \max_s \nu_s^{M_{max}} e^{\theta(RM_{max} - C + x)} \mathbb{P}(A(T_y) - CT_y \geq x) \\ &+ \max_s \nu_s^{M_{max}} e^{-\theta y} \ .\end{aligned}$$

Now simply let $y \to \infty$ and thus

$$\begin{aligned}\mathbb{E}[X_0] &\leq \max_s \nu_s^{M_{max}} e^{\theta(RM_{max} - C + x)} \mathbb{P}(T < \infty) \\ &= \frac{\mathbb{E}[X_0]}{H_l} e^{\theta x} \mathbb{P}(T < \infty)\end{aligned}$$

which completes the proof. □

C. Numerical Results

As in Section III, we next discuss the discrepancy between static and dynamic queues. Recall that the *exponential decay rate* θ from Theorem 2 is the same for the upper and lower bounds, respectively, and is thus the dominating factor for the decay of the overflow probability $\mathbb{P}(Q \geq x)$.

We consider a similar numerical settings as in Section III-E with an average $M_{avg} = 10$ of homogenous Markov-Modulated sources, as in Figure 4, which are active (i.e., dwelling in the states 0 and 1). Formally,

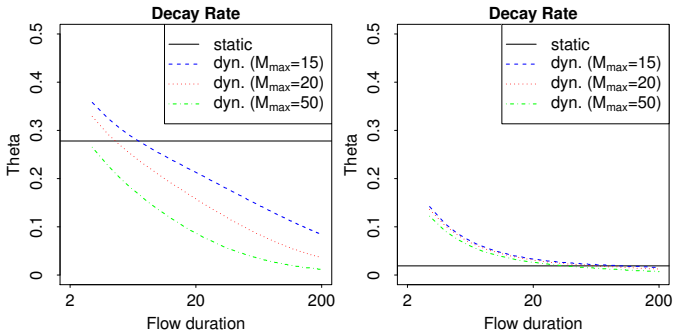
$$\pi_{\text{IA}} = \frac{r}{r + 1 - s} = 0.25 \ . \tag{15}$$

The parameter r determines the flow's average lifetime (which equals r^{-1}). Its range is the interval $[0, \frac{1}{3}]$; for $r = 0$ the queues are static, whereas for $r > \frac{1}{3}$ the parameter s cannot be scaled such that Eq. (15) holds. The ratio RC^{-1} is scaled such that the link utilization $\rho = 0.75$ remains constant in all cases, i.e.,

$$RC^{-1} = \frac{\rho}{\pi_1 M_{max}} \text{ and } RC^{-1} = \frac{\rho}{(\pi_{\mathbf{act}})_1 M_{avg}}$$

in the dynamic and static cases, respectively.

In Figure 5.(a) and (b) we illustrate the dominating factor θ from Theorem 2, of the probability of $\mathbb{P}(Q \geq x)$, for various average lifetimes r^{-1} of the flows. Compared to (a), the scenario from (b) captures burstier flows (by decreasing the transition probabilities by a factor of 10). In both figures we consider a static scenario (i.e., $M_{max} = 10$) and three (properly normalized) dynamic (dyn.) scenarios by varying $M_{max} = 15, 20, 50$.



(a) $p = 0.1$, $q = 0.5$ (b) $p = 0.01$, $q = 0.05$

Fig. 5. Decay rate θ as a function of the flows' average lifetime r^{-1} for both static and dynamic (dyn.) scenarios ($\rho = 0.75$, $M_{avg} = 10$, RC^{-1} is rescaled for each r^{-1}; the x-axis is shown on a log-scale)

Figure 5.(a) highlights the expected behavior that randomness in the number of flows 'hurts' the system's performance: Unless the flows are very short-lived (i.e., $r^{-1} \geq 5$) the backlog in the dynamic case is on average larger than its deterministic counterpart. Interestingly, for $r^{-1} \leq 4$ the performance actually benefits from randomization. This is due to the fact that for very short-lived flows, the (beneficial) property of multiplexing roughly *independent* flows (as the Markov structure lasts very shortly) outruns the (detrimental) effect of the *bursty* sources.

This phase-transition effect, i.e., the actual value of the flows' average lifetime at which dynamic multiplexing 'hurts', depends on the flows' own burstiness. This can be seen from Figure 5.(b) where the phase-transition occurs at much larger average lifetimes (and at which the flows remain roughly independent since the flows' Markov structure survives for around the average dwelling time in one of the states).

In conclusion, the figures indicate that for reasonable (i.e., not very short) average flows' lifetimes, flows' multiplexing 'hurts' the queue size. Moreover, the discrepancy between static and dynamic queues depends on the flows' own burstiness and also the distribution/support of the number of flows, and can be arbitrarily large as shown in Figure 5.(a) for large M_{max} and long flows.

V. Conclusions

In this paper we have investigated queueing behavior in typically neglected but highly relevant dynamic queues characterized by a *random number of parallel flows*. Under some strong iid assumptions, enabling a tractable analysis, we have first shown that dynamic queues retain some extremal properties from static queues, i.e., capacities are maximized by constant distributions and are minimized by bimodal distributions. While the iid case confirms that "*determinism minimizes the queues*", we have shown that this folk principle fails in the more realistic case when the number of parallel

flows has a Markov structure. Concretely, we have shown that there is a phase-transition of the flows' average lifetime, below which dynamic queues are smaller than static queues. While our observations jointly depend on the overall statistics, they nevertheless provide a convincing argument that current approximations of dynamic by static queues can be very misleading, and that a rigorous analysis of queueing scenarios with a dynamic number of flows is necessary.

ACKNOWLEDGEMENT

This work was partially funded by the DFG grant Ci 195/1-1.

REFERENCES

[1] J. Abate, G. L. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems*, 16(3-4):311–338, Sept. 1994.

[2] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical Journal*, 61(8):1871–1894, Oct. 1982.

[3] S. Asmussen, A. Frey, T. Rolski, and V. Schmidt. Does Markov-modulation increase the risk? *ASTIN Bulletin*, 25(1):49–66, May 1995.

[4] S. Asmussen and C. O'Cinneide. On the tail of the waiting time in a markov-modulated M/G/1 queue. *Operations Research*, 50(3):559–565, May-June 2002.

[5] N. Bansal. Analysis of the M/G/1 processor-sharing queue with bulk arrivals. *Operations Research Letters*, 31(5):401 – 405, Sept. 2003.

[6] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, Apr. 1975.

[7] N. Bäuerle and T. Rolski. A monotonicity result for the workload in Markov-modulated queues. *Journal of Applied Probability*, 35(3):741–747, Sept. 1998.

[8] D. Bertsimas, D. Gamarnik, and A. A. Rikun. Performance analysis of queueing networks via robust optimization. *Operations Research*, 59(2):455–466, Mar. 2011.

[9] P. Billingsley. *Probability and Measure (3^{rd} Edition)*. Wiley, 1995.

[10] T. Bonald and L. Massoulié. Impact of fairness on internet performance. In *ACM Sigmetrics*, pages 82–91, 2001.

[11] T. Bonald and A. Proutière. Flow-level stability of utility-based allocations for non-convex rate regions. In *40th Annual Conference on Information Sciences and Systems*, pages 327–332, 2006.

[12] P. J. Burke. Delays in single-server queues with batch input. *INFORMS-Operations Research*, 23(4):830–833, July-August 1975.

[13] A. Bušić, J.-M. Fourneau, and N. Pekergin. Worst case analysis of batch arrivals with the increasing convex ordering. In A. Horváth and M. Telek, editors, *Formal Methods and Stochastic Models for Performance Evaluation*, volume 4054 of *Lecture Notes in Computer Science*, pages 196–210. Springer, 2006.

[14] J. P. Buzen. Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM*, 16(9):527–531, Sept. 1973.

[15] C.-S. Chang. *Performance Guarantees in Communication Networks*. Springer Verlag, 2000.

[16] M. L. Chaudhry and J. G. C. Templeton. *A First Course in Bulk Queues*. John Wiley and Sons, 1983.

[17] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, Feb. 1996.

[18] F. Ciucu and J. Schmitt. Perspectives on network calculus - No free lunch but still good value. In *ACM Sigcomm*, 2012.

[19] N. G. Duffield. Exponential bounds for queues with Markovian arrivals. *Queueing Systems*, 17(3-4):413–430, Sept. 1994.

[20] S. B. Fred, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. In *ACM Sigcomm*, pages 111–122, 2001.

[21] B. Hajek. The proof of a folk theorem on queuing delay with applications to routing in networks. *Journal of the ACM*, 30(4):834–851, Oct. 1983.

[22] M. Harchol-Balter and D. Wolfe. Bounding delays in packet-routing networks. In *ACM Symposium on Theory of Computing (STOC)*, pages 248–257, 1995.

[23] P. A. Humblet. Determinism minimizes waiting time in queues. Technical report, MIT Laboratory for Information and Decision Systems, LIDS-P-1207, 1982.

[24] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, 2008.

[25] F. P. Kelly. Networks of queues with customers of different types. *Journal of Applied Probability*, 3(12):542–554, Sept. 1975.

[26] F. P. Kelly. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications. (Editors: F.P. Kelly, S. Zachary and I.B. Ziedins) Royal Statistical Society Lecture Notes Series, 4*, pages 141–168. Oxford University Press, 1996.

[27] J. F. C. Kingman. A martingale inequality in the theory of queues. *Cambridge Philosophical Society*, 60(2):359–361, Oct. 1964.

[28] D. C. Lee and J. N. Tsitsiklis. The worst bulk arrival process to a queue. Technical report, MIT Laboratory for Information and Decision Systems, LIDS-P-2116, 1992.

[29] J. Liu, A. Proutière, Y. Yi, M. Chiang, and H. Poor. Stability, fairness, and performance: A flow-level study on nonconvex and time-varying rate regions. *IEEE Transactions on Information Theory*, 55(8):3437–3456, Aug. 2009.

[30] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *Journal of the ACM*, 44(3):366–394, May 1997.

[31] A. M. Makowski and T. Philips. Simple proofs of some folk theorems for parallel queues. Technical report, Institute for Systems Research, ISR-TR-1989-37, 1989.

[32] L. Massoulié and J. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15(1-2):185–201, Nov. 2000.

[33] R. R. Mazumdar. *Performance Modeling, Loss Networks, and Statistical Multiplexing*. Synthesis Lectures on Communication Networks. Morgan & Claypool Publishers, 2009.

[34] M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queuing networks. *Journal of the ACM*, 27(2):313–322, Apr. 1980.

[35] B. Rogozin. Some extremal problems in the theory of mass service. *Theory of Probability & Its Applications*, 11(1):144–151, 1966.

[36] S. M. Ross. Bounds on the delay distribution in GI/G/1 queues. *Journal of Applied Probability*, 11(2):417–421, June 1974.

[37] S. M. Ross. Average delay in queues with non-stationary Poisson arrivals. *Journal of Applied Probability*, 15(3):602–609, Sept. 1978.

[38] M. Schleyer. An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals. *OR Spectrum*, 29(4):745–763, Oct. 2007.

[39] M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer, 2007.

[40] N. B. Shroff and M. Schwartz. Improved loss calculations at an ATM multiplexer. *IEEE/ACM Transactions on Networking*, 6(4):411–421, Aug. 1998.

[41] T. E. Stern and A. I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability*, 23(1):105–139, Mar. 1991.

[42] H. Takagi and Y. Takahashi. Priority queues with batch Poisson arrivals. *Operations Research Letters*, 10(4):225–232, June 1991.

[43] G. de Veciana, T.-J. Lee, and T. Konstantopoulos. Stability and performance analysis of networks supporting services with rate control - could the internet be unstable? In *IEEE Infocom*, pages 802–810, 1999.

[44] W. Whitt. The effect of variability in the GI/G/s queue. *Journal of Applied Probability*, 17(4):1062–1071, 1980.

[45] W. Whitt. Comparison conjectures about the M/G/s queue. *Operations Research Letters*, 2(5):203–210, Dec. 1983.

[46] W. Whitt. On approximations for queues, I: Extremal distributions. Technical report, Institute for Systems Research, ISR-TR-1989-37, 1989.

[47] D. D. W. Yao, M. L. Chaudhry, and J. G. C. Templeton. On bounds for bulk arrival queues. *European Journal of Operational Research*, 15(2):237 – 243, Feb. 1984.